



Entrez Program Utilities (EUtils)

Programmatic Access to NCBI Data

<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/>

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services



Scope and Access

The common way to access data stored at NCBI is through web browsers using the NCBI Entrez system, where they can be searched, retrieved for display and then downloaded in a selected format. For batch data access from the Entrez family of databases, bypassing the browser, NCBI offers Entrez Program Utilities (EUtils). This service offers esearch, esummary, efetch, elink, einfo and espell programs to provide search, document summary, full record retrieval, linking between records from the same or different databases, summary information for the databases and spelling correction services, respectively. This set of services is also available through Entrez Direct (eDirect), a package for Linux/Unix-based platform with command-only interface. Data from some databases may not be available for retrieval through the efetch function under EUtils. More detailed description of EUtils is at www.ncbi.nlm.nih.gov/books/NBK25501/. The document for eDirect package is at www.ncbi.nlm.nih.gov/books/NBK179288/.

A Perl-based program suite, named Ebot (www.ncbi.nlm.nih.gov/Class/PowerTools/eutils/ebot/ebot.cgi), can be used to create custom scripts for accessing data through the EUtils service. Other NCBI databases and services also offer programmatic access outside the EUtils framework. Some will be mentioned below.

Esearch: searching a database with query terms to retrieve a list of uids

Esearch is the counterpart to web search. It takes a set of query terms, in URL encoded form, and searches against the selected database. Esearch returns the search result in XML format. The result can contain a reference to a search history in the form of a *WebEnv* and *QueryKey* if *usehistory=y* is used. This reference can be used in subsequent operations in Esearch or other EUtils operations. Esearch requires db and term.

Parameter	Function and format
db	Database name, e.g., db=nuccore and db=pubmed
term	Query text string with escaped white spaces and other non-
usehistory	Use usehistory=y to reference a previous WebEnv or to ask for WebEnv and QueryKey instead of a list of uids
WebEnv	Text string referencing a set of cached results

The following URL searches PubMed records

pertaining to the terms “Neanderthal genome sequence”:

<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=neanderthal+genome+sequence>

```
<eSearchResult><Count>19</Count><RetMax>18</RetMax><RetStart>0</RetStart>
<IdList>
    <Id>21453001</Id>
    ...
</IdList>
...
</eSearchResult>
```

(Sample excerpt for the above request)

To search for mRNA Reference Sequences and request a reference of search history instead, the following URL can be used, with the reference represented by the WebEnv and QueryKey fields in the result:

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=nuccore&term=refseq%5Bfilter%5D+AND+biomol_mrna%5Bprop%5D&usehistory=y

```
<QueryKey>1</QueryKey>
<WebEnv>NCID_1_32921916_130.14.18.47_9001_1300717375_377448279</WebEnv>
```

(Example QueryKey and WebEnv)

The QueryKey value from a previous Esearch can be used as part of a new Esearch request. For the above request, the search result can be represented by that number in #1 format (with # escaped to URL-safe format) and combined with new query terms to get a more specific subset. More details are given in the EUtils help manual.

Note: There are pending changes in how sequence databases operate. Two webinar recordings on this topic are available from NCBI's YouTube channel: <https://youtu.be/rIDQEnnOr6g> and https://youtu.be/O2_n6cBuFwE.

Esummary: retrieving the document summary for a list of uids

Esummary takes a list of uids and retrieves the document summary for those records from the target database as specified by the input uids. Esummary returns the results in XML format. The fields in the results vary according to the target database.

The following URL retrieves the document summary for pubmed record 9499025 and 11395195:

<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=pubmed&id=9499025,11395195>

Esummary can also take a WebEnv and QueryKey pair returned by Esearch or Elink as inputs instead of a list of uids, as in the example URL below:

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esummary.fcgi?db=pubmed&query_key=1&WebEnv=NCID_1_27748375_130.14.22.148_9001_1300480201_685518290

For most databases, Esummary returns more information than its web Entrez counterpart. This reduces the need to fetch the complete record, which is often more computationally costly. An excerpt from the first Esummary call is given below.

```
<eSummaryResult>
  <DocSum>
    <Id>9499025</Id>
    <Item Name="PubDate" Type="Date">1998 Mar</Item>
    <Item Name="EPubDate" Type="Date" />
    <Item Name="Source" Type="String">J Virol</Item>
    <Item Name="AuthorList" Type="List">
      <Item Name="Author" Type="String">Skiadopoulos MH</Item>
      ...
    </Item>
    <Item Name="LastAuthor" Type="String">Murphy BR</Item>
    <Item Name="Title" Type="String">Three amino acid substitutions in the L protein of the human parainfluenza virus type 3 cp45 live attenuated vaccine candidate contribute to its temperature-sensitive and attenuation phenotypes.</Item>
    <Item Name="Volume" Type="String">72</Item>
    <Item Name="Issue" Type="String">3</Item>
    <Item Name="Pages" Type="String">1762-8</Item>
    ...
  </DocSum>
  ...
</eSummaryResult>
```

(An excerpt of a PubMed esummary result)

Efetch: takes a list of uids and retrieves the records in specified formats

Efetch is needed to retrieve a database record in its complete form or in a specific display format other than uid or document summary. For example, for evaluation of sequence similarities, the FASTA sequence of a sequence record will be needed, while for importation of a PubMed abstract to third party tools, the Medline format may be required. Efatch takes a list of comma-separated uids and a target database as input and returns the complete records in ASN.1 format or a specific format as specified by the combination of settings of rettype and retmode. Representative settings for commonly requested retrieval formats for PubMed, Nucleotide, Protein and Gene databases are given in the table above.

Retrieval format	Combination of settings	
	rettype	retmode
PubMed abstract in XML format	rettype=xml	retmode=text
PubMed abstract in Medline format	rettype=medline	retmode=text
Nucleotide sequence in gb xml	rettype=xml	retmode=text
Nucleotide sequence in FASTA text	rettype=fasta	retmode=text
Protein sequence in genpept format	rettype=gp	retmode=text
Gene record in full xml format	rettype=xml	retmode=text

It should be noted that Efatch support is not uniform across Entrez databases. Only a subset of databases provides full Efatch support. The Efatch support for some Entrez databases, such as SNP, GEO and Taxonomy, deviates from those described above. Efatch specifics for these databases are available from their homepages. The PubChem family of databases (pccompound, pcs substance and pcassay) provides a fetch service through its own PUG setup as described at <https://pubchem.ncbi.nlm.nih.gov/pug/pughelp.html>.

Elink: retrieving related uids from a target database for a set of input uids

The power of the NCBI Entrez system is its capability to organize the database records from the same or different databases using pre-calculated links. In web Entrez, the related information appears as items under the “All links from this record” column. The same information is also available through the Elink program under EUtils. Specifically, Elink takes a uid from a source database and retrieves the uids for the related records from the target database. The example below retrieves the related articles for an article in PubMed using its pmid 11395195:

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=11395195&db=pubmed&linkname=pubmed_pubmed

```
<eLinkResult>
  <LinkSet>
    <DbFrom>pubmed</DbFrom>
      <IdList>
        <Id>11395195</Id>
      </IdList>
      <LinkSetDb>
        <DbTo>pubmed</DbTo>
        <LinkName>pubmed_pubmed</LinkName>
        <Link><Id>11395195</Id></Link>
        <Link><Id>10864657</Id></Link>
        ...
      </LinkSetDb>
    </LinkSet>
  </eLinkResult>
```

(An excerpt of the elink result from the above URL)

For PubMed records with full text, the publisher or full-text provider link can be obtained using Elink as in:

<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=11395195&cmd=prlinks>

Einfo: returning a list of fields available for a supported database

Knowing fields available for a target database will help the construction of rational and efficient EUtils request URLs. Einfo is the tool that does just that. The Einfo program functions in two modes. Without the db specification, it returns a list of supported databases. The database name can be used to retrieve more detailed information pertaining to that database.

<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi>

<https://eutils.ncbi.nlm.nih.gov/entrez/eutils/einfo.fcgi?db=pubmed>

```
<eInfoResult>
  <DbInfo>
    <DbName>pubmed</DbName>
    <MenuName>PubMed</MenuName>
      <Description>PubMed bibliographic record</Description>
      <Count>21382264</Count>
      <LastUpdate>2011/12/11 05:12</LastUpdate>
      <FieldList>
        <Field>
          <Name>ALL</Name>
          <FullName>All Fields</FullName>
          ...
        </Field>
      </FieldList>
      <LinkList>
        <Link>
          <Name>pubmed_biosample</Name>
          <Menu>BioSample Links</Menu>
          <Description>BioSample links</Description>
          <DbTo>biosample</DbTo>
        </Link>
        ...
        <Link>
          <Name>pubmed_pmc</Name>
          <Menu>PMC Links</Menu>
          <Description>Free full-text versions of the current articles in the PubMed Central database.</Description>
          <DbTo>pmc</DbTo>
        </Link>
      </LinkList>
    </DbInfo>
  </eInfoResult>
```

(An excerpt of the einfo result for the PubMed database)

Putting Things Together: linking different EUtils calls using WebEnv

Using the WebEnv and QueryKey, it is possible to connect individual calls to different EUtils services into an organized workflow. For example, to retrieve the FASTA sequences for a set of RefSeq protein records pertaining to a list of human Genes of interest, one needs to do:

- Elink from gene with a list of geneid's to protein database with specific linkname
- Efetch from the protein database referencing the WebEnv and QueryKey to get the FASTA sequence

The first step can be accomplished using this Elink URL: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=gene&db=protein&id=3077&cmd=neighbor_history

In the result, two types of links are found, both referencing the same WebEnv, but pointing to different subsets with different QueryKey values. The one needed has a QueryKey value of 2.

Using the QueryKey and WebEnv, the FASTA sequences of the RefSeq proteins can be retrieved using efetch:

```
<eLinkResult>
<LinkSet>
  <DbFrom>gene</DbFrom>
  <IdList>
    <Id>3077</Id>
  </IdList>
  <LinkSetDbHistory>
    <DbTo>protein</DbTo>
    <LinkName>gene_protein</LinkName>
    <QueryKey>1</QueryKey>
  </LinkSetDbHistory>
  <LinkSetDbHistory>
    <DbTo>protein</DbTo>
    <LinkName>gene_protein_refseq</LinkName>
    <QueryKey>2</QueryKey>
  </LinkSetDbHistory>
  <WebEnv>NCID_1_33866643_130.14.22.148_9001_1300741898_1456502256</WebEnv>
</LinkSet>
</eLinkResult>
```

(Elink output from geneid 3077 to protein)

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=protein&rettype=fasta&retmode=text&query_key=2&WebEnv=NCID_1_33866643_130.14.22.148_9001_1300741898_1456502256

```
>gi|21040357|ref|NP_620580.1| hereditary hemochromatosis protein isoform 11 precursor [Homo sapiens]
MGPRARPALLLMLLQTAVLQGRLLQPSPSGTTLVIGVISGIAVFVVILFIGILFIILRKQRGSRGAMGHY
VIAERE

>gi|21040355|ref|NP_620579.1| hereditary hemochromatosis protein isoform 10 precursor [Homo sapiens]
MGPRARPALLLMLLQTAVLQGRLLLPPPLVKVTHTVTSSVTLRCRALNYPQNITMKWLKDQPMDAKE
FEPKDVLPNGDGTYQGWITLAVERPGEERQRTCTQVEHPGLDQPLIVIWEPSPSGTLVIGVISGIAVFVVIL
FIGILFIILRKQRGSRGAMGHYVIAERE
.... . . .
```

The WebEnv and QueryKey can also be used to retrieve the uids (GIs) for this set of protein records:

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=protein&query_key=2&WebEnv=NCID_1_33866643_130.14.22.148_9001_1300741898_1456502256

The retrieved GIs can be parsed out programmatically and used as input to other NCBI services, such as BLAST through BLAST URLAPI (<https://www.ncbi.nlm.nih.gov/blast/Doc/urlapi.html>).

```
<eSearchResult>
<Count>9</Count><RetMax>9</RetMax><RetStart>0</RetStart>
<IdList>
  <Id>21040357</Id>
  <Id>21040355</Id>
  <Id>21040353</Id>
  <Id>21040351</Id>
  ...
</IdList><TranslationSet/><TranslationStack><OP>GROUP</OP> </TranslationStack><QueryTranslation>#2</
QueryTranslation>
</eSearchResult>
```

Please subscribe to the Utilities announcement list to get notified on pending update and other changes:
<https://www.ncbi.nlm.nih.gov/mailman/listinfo/utilities-announce>